MAYO CLINIC PROCEEDINGS: DIGITAL HEALTH

Artificial Intelligence for Kidney Stone Spectra Analysis: Using Artificial Intelligence Algorithms for Quality Assurance in the Clinical Laboratory

Patrick L. Day, MPH; Sarah Erdahl, MS; Denise L. Rokke, BS; Mikolaj Wieczorek, BS; Patrick W. Johnson, BS; Paul J. Jannetto, PhD; Joshua A. Bornhorst, PhD; and Rickey E. Carter, PhD

Abstract

Objective: To determine if a set of artificial intelligence (AI) algorithms could be leveraged to interpret Fourier transform infrared spectroscopy (FTIR) spectra and detect potentially erroneous stone composition results reported in the laboratory information system by the clinical laboratory.

Background: Nephrolithiasis (kidney stones) is highly prevalent, causes significant pain, and costs billions of dollars annually to treat and prevent. Currently, FTIR is considered the reference method for clinical kidney stone constituent analysis. This process, however, involves human interpretation of spectra by a qualified technologist and is susceptible to human error.

Methods: This prospective validation study was conducted from October 29, 2020, to October 28, 2021, to test if the addition of AI algorithm overreads to FTIR spectra could improve the detection rate of technologist-misclassified FTIR spectra. The preceding year was used as a control period. Disagreement between the AI overread and technician interpretation was resolved by an independent human interpretation. The rate of verified human misclassifications that resulted in revised reported results was the primary end point.

Results: Spectra of 81,517 kidney stones were reviewed over the course of 1 year. The overall clinical concordance between the technologist and algorithm was 90.0% (73,388/81,517). The report revision rate during the AI implementation period was nearly 8 times higher than that during the control period (relative risk, 7.9; 95% CI, 4.1−15.2).

Conclusion: This study demonstrated that an AI quality assurance check of human spectra interpretation resulted in the identification of a significant increase in erroneously classified spectra by clinical laboratory technologists.

■ Mayo Clin Proc Digital Health 2023;1(1):1-12

Nephrolithiasis (kidney stones) is highly prevalent (∼12%) worldwide, and its incidence has increased over the past several decades.[1,2] Kidney stones cause approximately 1.3 million emergency department visits annually and result in an annual economic burden of >5 billion USD.[3,4] Additionally, kidney stones are extremely painful for the patient. Although it can often be prevented, kidney stone recurrence is common.[5]

To effectively treat and prevent future kidney stones, accurate identification of kidney stone constituents is often necessary to guide proper treatment.[6] For example, if calcium oxalate is identified as the major stone constituent, citric acid supplementation may be recommended. If the patient exhibited an associated urinary risk of high uric acid, then diet modifications and/or use of allopurinol might be prescribed. If calcium phosphate is identified in kidney stones of women of

From the Department of Laboratory Medicine and Pathology (P.L.D., S.E., D.L.R., J.A.B.), Mayo Clinic, Rochester, MN; and Department of Quantitative Health Sciences (M.W., P.W.J., R.E.C.) and Digital Innovation Lab (M.W., P.W.J., R.E.C.), Mayo Clinic, Jacksonville, FL.

childbearing age, a pregnancy test may be recommended. If uric acid is identified within the kidney stone, treatment to alkalize the urine may be recommended. Additional treatments related to uric acid identification in stones are increase in fluid intake, decrease in dietary protein, and possible use of potassium citrate. Alternatively, if struvite is identified in the kidney stone, the treatment recommended would be to acidify the urine.[7] As is apparent in these situations, treatment can differ significantly depending on the constituents of the kidney stone. Therefore, if a kidney stone is falsely characterized, the patient could be potentially treated incorrectly, which could promote the occurrence of future kidney stones or complicate the treatment of existing kidney stones. However, the possible severity of kidney stone misclassification is dependent on the individual discrepancy scenario.

Fourier transform infrared spectroscopy (FTIR) is considered the reference method for clinical kidney stone constituent analysis.[6,8] Using FTIR, kidney stone constituents can be identified by percent mass present in the stone sample. However, to date, FTIR relies on a highly trained technologist to visually interpret the kidney stone spectrum and report the results into the laboratory information system (LIS). Additionally, as part of good clinical practice and regulatory adherence, a secondary review by another technologist is required to verify the manually entered results before releasing the results to the ordering physician. This standard FTIR workflow is labor intensive, time-consuming, and susceptible to human errors, such as spectra misinterpretation, typographical errors when transcribing results into the LIS, and failure to notice reporting errors during secondary review.

Advances in health care artificial intelligence (AI) have allowed for quantifiable improvements in clinical reporting, clinical alert systems, and adverse drug reaction monitoring, all of which improve quality and patient safety.[9] AI is currently being used in the clinical laboratory to diagnose disease from images and automatically release laboratory results to the ordering physician.[10] AI image recognition has immense potential to advance digital pathology and laboratory medicine as it relates to improved diagnosis, reduced labor, and improved turnaround times.[11] With all these potential benefits, the primary goal of this prospective quality assurance program was to determine if an internally developed set of AI algorithms could be leveraged to interpret kidney stone FTIR spectra after the technicians reported their findings to detect potentially erroneous stone composition results reported in the LIS.

METHODS

A detailed supplemental appendix provides complete methodological details, particularly for the development and validation of the machine learning approaches. This study was conducted in 2 phases. The first phase included the development of AI algorithms using spectra readouts from FTIR instruments and technologist-derived interpretations of the spectra. The second phase was a prospective quality assurance project in which the spectra were analyzed by the algorithms with reports provided to the technologists. No human subject contact was made, and no identifiable data were available in either of these studies. As such, the research did not constitute human subject research.

FTIR Analysis of Kidney Stones

Physical homogenization was used to create a fine powder of dry kidney stones. A portion of this fine powder was transferred to a PerkinElmer Frontier FTIR spectrometer with a Universal ATR diamond/ZnSe crystal sampling accessory. An FTIR spectra was then obtained. Each of the FTIR spectra was evaluated by a highly trained technologist and confirmed by a second technologist using the FTIR software that includes a laboratory search function and commercial spectral libraries. It provides a list of spectral matches and match scores. Other functions such as spectral overlay or subtraction are available to the technologist to assist in the interpretation of the spectra. Before releasing any clinical results, each technologist was fully trained, and competency in interpreting kidney stone spectra and using the electronic kidney stone spectra libraries was documented. The technologists used internally developed guidelines using specific FTIR wavelengths to minimize variability in interpretation calls associated with multiple technologists interpreting kidney stone spectra.

Algorithm Development

The Metals Laboratory at Mayo Clinic has analyzed >1,000,000 kidney stones using this clinically-validated FTIR process. This has resulted in a large database of clinical kidney stone FTIR spectra with associated stone composition information—an ideal setting to develop AI models to provide clinical decision support to the laboratory practices. An additional resource is an internally developed kidney stone FTIR spectra library with >300 different stone composition spectra. Both the historical clinical kidney stone FTIR spectra database and internal kidney stone library spectra were used for algorithm training and validation. A detailed description of the development and validation of the AI algorithms can be found in the supplemental material section of this article.

Quality Assurance

Using these internally developed AI algorithms to interpret the FTIR spectra, a quality assurance report was generated twice a week to analyze clinical kidney stone spectra and compare these interpretations to technologist interpretations in the LIS. The quality assurance program generated a list of stone spectra with discordance between the report by the technologist in the LIS and the AI program interpretation for the same kidney stone spectra. Incongruent spectra interpretations were then reviewed by a qualified technologist to determine if an incorrect result was reported in the LIS by the technologist.

For this study, the quality assurance program reviewed kidney stone spectra that were generated and reported in the LIS between October 29, 2020, and October 28, 2021. The number of incorrect results reported in the LIS would be compared to the number of kidney stone laboratory events 1 year before initiating the AI quality assurance program to determine the number of additional incorrect kidney stones results that could be identified and corrected with the incorporation of this AI quality assurance program.

Statistical Considerations

Standard descriptive measures of diagnostic performance were included in the data summarization. Real-world model accuracy was defined in 2 ways. Concordance was defined as the AI algorithm prediction exactly matching the kidney stone results in the LIS report. However, during the standard clinical workflow, the initial spectra interpretation of kidney stones was reviewed by a second technologist before being released to the physician. Thus, some variability in results is expected. The established laboratory guidelines (not associated with this study or newly defined for the interpretation of the AI predictions) include provisions that allow variability of the percent composition of the individual constituents (±20%), provided all constituents present are identified in the prediction. Our second concordance metric applied the same clinical guidelines to the agreement between the AI algorithms and final clinical result reported into the LIS. For example, a stone with 80% calcium oxalate monohydrate and 20% calcium oxalate dihydrate would be considered a correct prediction if the AI algorithms reported the composition of these 2 constituents as any of the following: 60%:40%, 70%:30%, 80%:20%, or 90%:10%. A prediction of 100% calcium oxalate monohydrate would not be considered a comparable classification because the dihydrate component was not identified. Using these established and well-defined guidelines, the primary concordance measurement (concordance pass QA) was tabulated for each comparison between AI and the technologist.

The revision rate in human stone classification during the quality assurance period was compared with the revision rate in the previous year using the rate ratio and its 95% CI. The revision rate ratio was calculated separately for the overall revision rate and for revisions that were motivated by internal staff and referring providers. The area under the receiver operating characteristic curve along with measures of diagnostic performance using a classification threshold of 0.5 was utilized. For this analysis, the sensitivity estimate was interpretable as the detection of a particular constituent in any percentage in the spectra.

To better understand the relationship of the convolutional neural network and its associated predictions, SHapley Additive exPlanations (SHAP)[12] were generated to verify the algorithm-learned features aligned with the knowledge of how the individual IR wavelengths should be modified by the individual constituents. SHAP was computed using the SHAP package version 0.39.0 for Python 3.8.6. Statistical analysis was conducted using R version 4.0.3.

RESULTS

Before the formal implementation of the algorithms into practice in support of the yearlong quality assurance program, an extensive model development and validation activity was conducted to develop an augmented workflow to support the integration of AI algorithms into the FTIR spectra classification process (Figure 1). Supplemental Table 1 (available online at https://www.mcpdigitalhealth.org/) summarizes the overall model concordance between the AI prediction and technician read according to stone classification on the 16,491 spectra randomly selected to be a validation sample during the algorithm development phase of the process. The algorithm was designed to provide a set of rules that provided a means to allow the FTIR spectra to "autoPass" human review in future implementations based on metrics of model confidence and relative match with a library of reference spectra (Supplemental Table 2, available online at https://www.mcpdigitalhealth.org/). The algorithm concordance was 95.2% (95% CI, 94.9%-95.5%) over 16,491 stones classified into 708 unique stone types. Approximately 50% (7,322/16,491) of stones met the criteria for autoPass, and the concordance with the technician was 99.5% (95% CI, 99.3%-99.7%). The concordance was 91.7% (95%, 91.2%-92.3%) in the remaining stones that would be candidates for human review.

The concordances between the AI predictions and technician results were high for the 2 most prevalent stone types, namely, 100% calcium oxalate monohydrate (4808/16,491; 29.2%) and 100% uric acid (974/16,491; 5.9%) ([99.5%; 95% CI, 99.3%-99.7%] and [99.6%; 95% CI, 99.1%-99.9%], respectively). Using the more stringent criteria to allow for autoPass, the accuracy of the predictions was 100% for both stone types. Importantly, for these 2 stone types, >88% of all stones were candidates for autoPass. Supplemental Table 1 lists the detailed performance of the 18 additional stone types. The concordance rates for these stones were all >90%. However, the percentage of stones meeting the autoPass criteria was lower.

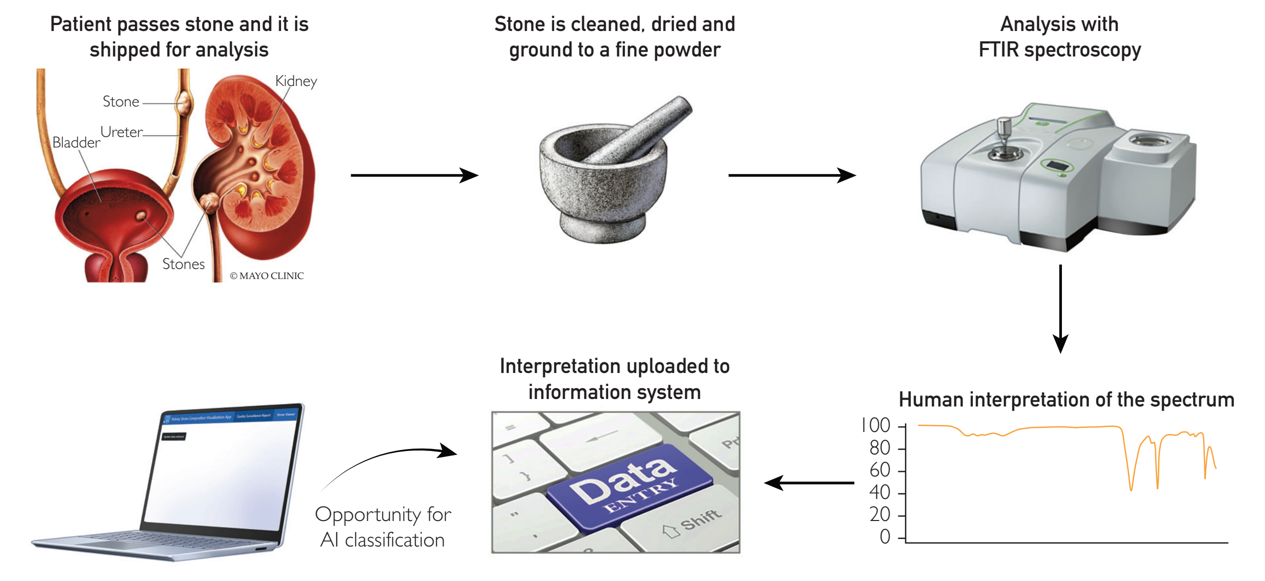
Patient passes stone and it is shipped for analysis
Stone
Kidney
Ureter
Bladder
Stones
© MAYO CLINIC
Stone is cleaned, dried and ground to a fine powder
Analysis with FTIR spectroscopy
Human interpretation of the spectrum
100
80
60
40
20
0
Interpretation uploaded to information system
Data
ENTRY
Shift
Opportunity for AI classification

FIGURE 1. Workflow of Fourier Transformed Infrared (FTIR) spectra classification process

Although the model performance was promising as the algorithms were being developed and evaluated in terms of concordance rate, there was a need to investigate the network's performance in the context of explainable AI.[13] For this purpose, we sought to examine how the model's predictions were formed relative to how we know human classification is performed on the spectra, namely, the visual inspection of areas of light absorption ("peaks" in the spectra). To validate the model predictions using a framework of explainable AI, SHAP[12] predictions across a range of correct and incorrect predictions were generated. Examples of SHAP predictions are illustrated in Figure 2. The heatmap intensities represent the relative weights of spectra locations that contributed to the model predictions. Areas of increased heatmap intensity were associated with unique wavelength regions for specific stones (Figure 2). Qualified technologists reviewed the heatmaps for these specific stones (struvite and ammonium urate) and concluded that the areas of increased heatmap intensity were associated with specific wavelength regions that are used to identify these kidney stone types from other commonly encountered kidney stones.

Prospective Quality Assurance Program Results

After model validation using previously collected FTIR spectra was complete, the prospective quality assurance program was initiated to measure the real-world performance of the algorithms along with the potential impact on the practice. The present study reviewed 81,517 kidney stone spectra as a part of the program that ran from October 29, 2020, to October 28, 2021. The AI algorithm and technologist provided concordant interpretations on 73,388 (90.0%; 95% CI, 89.8%-90.2%) clinical kidney stone spectra (Table 1). When the stone spectra were grouped into those stones selected for autoPass (candidate for automatic classification) and those not selected (flagged for manual secondary review), the concordance rates were 98.7% (95% CI, 98.56%-98.9%) and 84.3% (95% CI, 84.0%-84.6%), respectively. Receiver operating characteristic curves were created for the major stone constituents to determine the sensitivity and specificity of the algorithm to correctly identify the presence or absence of specific stone constituents in certain stone samples. Table 2 summarizes diagnostic performance metrics for calcium oxalate monohydrate, calcium oxalate dihydrate, calcium phosphate (apatite), and uric acid, 4 commonly identified stone constituents. Calcium oxalate monohydrate was present in 73.1% (59,581/81,517) of all study stones, calcium phosphate was present in 45.5% (37,080/81,517) of stones, calcium oxalate dihydrate was present in 38.5% (31,391/81,517) of stones, and uric acid was present in 10.3% (8,366/81,517) of stones. Further details on model performance on these constituents is provided in Table 3.

Review of the discordant spectra demonstrated that 62 stone spectra were incorrectly reported in the LIS by the laboratory but subsequently identified by the AI quality assurance program during the study period. These misreported stones would not have been discovered without the AI quality assurance program. These misreported stones were subsequently entered into the Laboratory Event Management System (LEMS), and the results were amended in the LIS and relayed to the ordering provider usually within 1 day of discovering the error. The laboratory then compared the number of misreported stones per 10,000 spectra in our study period with the number of misreported stones per 10,000 spectra 1 year before starting the kidney stone quality assurance program (Table 4). Three main reasons exist for the amended kidney stone results in the LIS. First, the laboratory observed and corrected a result entered into the LIS during normal clinical operation. Second, the treating physician called to either confirm or clarify the result. Finally, the error was identified by the new AI quality assurance program. As is shown, the overall rate of misreported kidney stone spectra in the study period was nearly 8 times higher than that in the 1 year before initiating the program (RR, 7.87; 95% CI, 4.08-15.2). This increase was clearly attributable to the implementation of the AI quality assurance program as the rate of misreported spectra identified by AI was 7.61 per 10,000 spectra during the study period. Although AI implementation was a designed change that could explain the overall increase in the relative

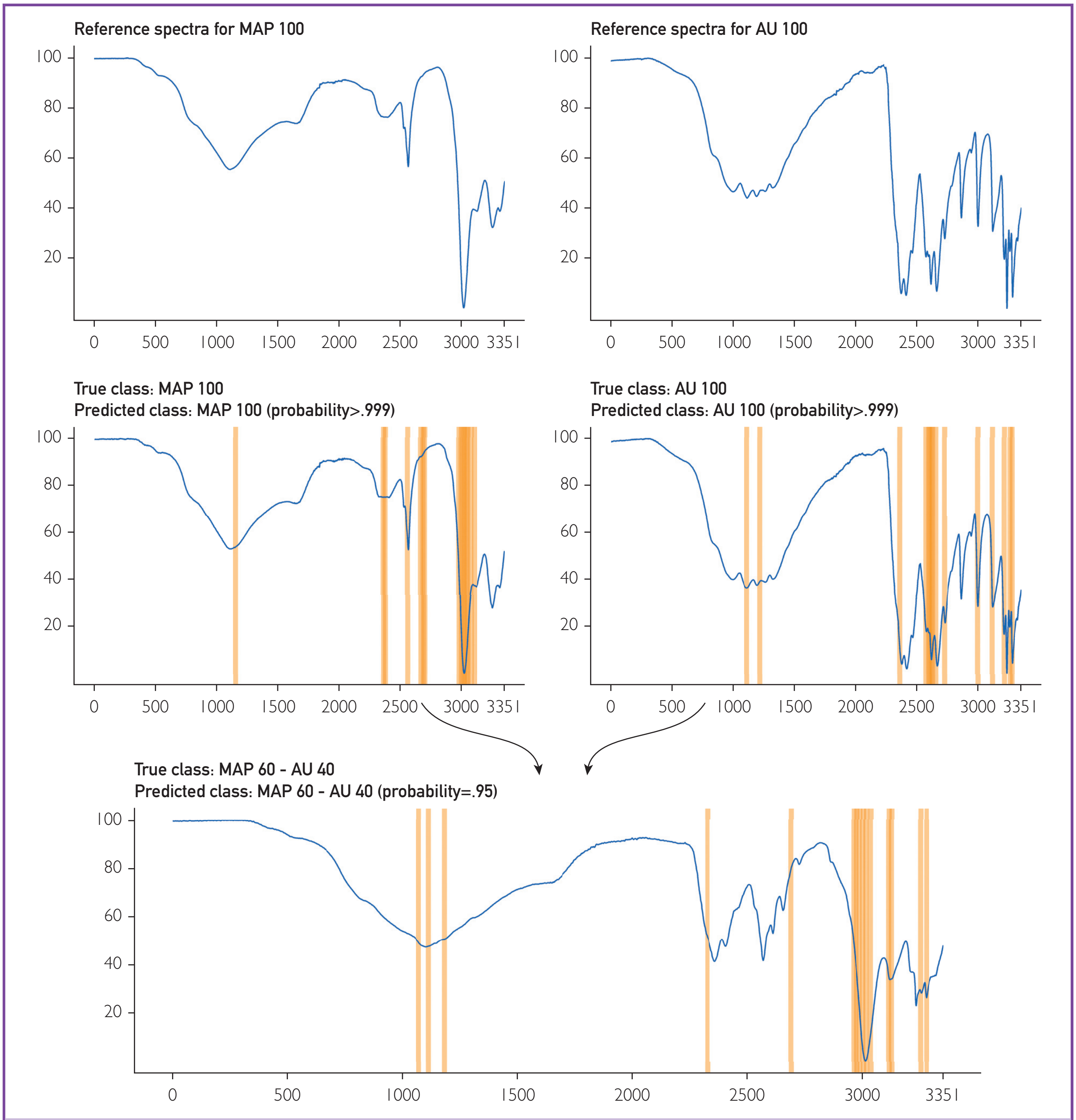

FIGURE 2. Example of the explainability of the model predictions using SHAP activation profiles. The first row illustrates 2 different reference spectra, pure magnesium ammonium phosphate (MAP, struvite) and ammonium urate (AU). The second row presents the SHAP activation profiles that contributed to the prediction with high confidence (>0.999 for both spectra). SHAP values were generated for the initial validation samples using 600 randomly selected training samples as the background. Fifteen IF features with the largest absolute SHAP values were selected and overlayed on the spectra signal identifying most important individual wavelengths to the prediction. The third row illustrates how the model predictions function on a spectra that combines the MAP and AU in a 60:40 ratio. The activation profiles for this combined element demonstrate how features from the pure spectra are combined into this mixed stone type.

TABLE 1. Combined "Overall" Model Performance with the 20 Most Common Stone Classifications for Prospective QA Program Results (n=81,517)

	Model accuracy		autoPass accuracy	
Stone constituent	Concordance rate (95% CI)	Concordance rate pass QA (95% CI)	% autoPass	Concordance rate pass QA (95% CI)
Overall performance				
All stones	67.3% (67.0%-67.6%) 54,876/81,517	90.0% (89.8%-90.2%) 73,388/81,517	-	-
autoPass	95.3% (95.1%-95.6%) 30,767/32,273	98.7% (98.6%-98.9%) 31,865/32,273	-	-
Manual review	49.0% (48.5%-49.4%) 24,109/49,244	84.3% (84.0%-84.6%) 41,523/49,244	-	-
Stone performance				
COM 100	89.3% (88.9%-89.6%) 23,926/26,801	98.1% (98.0%-98.3%) 26,304/26,801	80.2% (21,489/26,801)	99.7% (99.6%-99.8%) 21,420/21,489
UA 100	98.0% (97.5%-98.3%) 4,410/4,502	99.5% (99.3%-99.7%) 4,480/4,502	87.6% (3,943/4,502)	100.0% (99.9%-100.0%) 3,943/3,943
COM 90—APA 10	74.8% (73.1%-76.4%) 2,127/2,845	97.5% (96.9%-98.1%) 2,775/2,845	54.2% (1,543/2,845)	99.6% (99.2%-99.9%) 1,537/1,543
COM 80—APA 20	59.9% (57.8%-62.0%) 1,304/2,176	97.1% (96.3%-97.8%) 2,113/2,176	11.4% (249/2,176)	97.6% (94.8%-99.1%) 243/249
COM 90—COD 10	82.1% (80.4%-83.7%) 1,702/2,073	98.1% (97.4%-98.6%) 2,033/2,073	46.1% (955/2,073)	100.0% (99.6%-100.0%) 955/955
APA 100	80.1% (77.9%-82.2%) 1,127/1,407	86.5% (84.6%-88.2%) 1,217/1,407	49.3% (694/1,407)	97.6% (96.1%-98.6%) 677/694
COD 100	79.5% (77.3%-81.6%) 1,089/1,370	96.9% (95.9%-97.8%) 1,328/1,370	55.7% (763/1,370)	100.0% (99.5%-100.0%) 763/763
COM 80—COD 20	73.1% (70.5%-75.6%) 885/1,210	89.2% (87.3%-90.9%) 1,079/1,210	69.9% (846/1,210)	99.2% (98.3%-99.7%) 839/846
APA 90—COD 10	70.8% (67.8%-73.6%) 698/986	91.4% (89.5%-93.1%) 901/986	40.6% (400/986)	99.5% (98.2%-99.9%) 398/400
COM 70—APA 30	46.1% (42.9%-49.2%) 451/979	96.5% (95.2%-97.6%) 945/979	1.1% (11/979)	90.9% (58.7%-99.8%) 10/11
COM 60—APA 30 - COD 10	59.1% (55.7%-62.4%) 506/856	98.1% (97.0%-98.9%) 840/856	0.0% (0/856)	-
COM 70—APA 20 - COD 10	59.1% (55.7%-62.4%) 506/856	98.5% (97.4%-99.2%) 843/856	1.4% (12/856)	100.0% (73.5%-100.0%) 12/12
COM 80—COD 10 - APA 10	70.5% (67.3%-73.5%) 601/853	94.5% (92.7%-95.9%) 806/853	20.0% (171/853)	98.2% (95.0%-99.6%) 168/171
APA 60—COD 20 - COM 20	54.0% (50.5%-57.5%) 438/811	96.2% (94.6%-97.4%) 780/811	0.0% (0/811)	-
COM 50—APA 40 - COD 10	38.6% (35.1%-42.1%) 296/767	97.8% (96.5%-98.7%) 750/767	0.1% (1/767)	0.0% (0.0%-97.5%) 0/1
COM 60—COD 20 - APA 20	63.4% (59.9%-66.8%) 485/765	98.2% (96.9%-99.0%) 751/765	0.3% (2/765)	50.0% (1.3%-98.7%) 1/2
APA 80—COD 10 - COM 10	53.9% (50.3%-57.5%) 410/761	69.4% (66.0%-72.6%) 528/761	13.8% (105/761)	55.2% (45.2%-65.0%) 58/105
APA 60—COM 30 - COD 10	65.9% (62.4%-69.3%) 501/760	96.7% (95.2%-97.9%) 735/760	0.0% (0/760)	-
COM 70—COD 20 - APA 10	67.5% (63.9%-70.9%) 480/711	94.4% (92.4%-96.0%) 671/711	1.0% (7/711)	14.3% (0.4%-57.9%) 1/7
APA 70—COM 20 - COD 10	56.8% (53.1%-60.5%) 399/702	93.3% (91.2%-95.0%) 655/702	0.7% (5/702)	80.0% (28.4%-99.5%) 4/5

Concordance = number of true classifications (exact matches), Concordance rate pass QA = number of true classifications after application of quality assurance rules (comparable matches), Concordance rate = number of true classifications/total spectra reviewed.

APA, calcium phosphate (apatite); autoPass, automatic classification; COD, calcium oxalate dihydrate; COM, calcium oxalate monohydrate; MAP, magnesium ammonium phosphate (struvite); UA, uric acid.

TABLE 2. Diagnostic Performance Measures by 4 Most Prevalent Stone Types for Prospective QA Program Results (n=81,517)

Measure (95% CI)	Calcium oxalate monohydrate	Calcium oxalate dihydrate	Calcium phosphate	Uric acid
Area under the Curve	0.995 (0.994-0.995)	0.985 (0.985-0.986)	0.994 (0.994-0.995)	0.997 (0.997-0.998)
Accuracy	97.3% (97.2%-97.4%) 79,333/81,517	92.4% (92.2%-92.6%) 75,328/81,517	96.2% (96.1%-96.4%) 78,441/81,517	99.2% (99.1%-99.3%) 80,867/81,517
Sensitivity (recall)	98.2% (98.1%-98.3%) 58,872/59,967	96.8% (96.6%-97.0%) 30,663/31,667	95.0% (94.7%-95.2%) 35,638/37,529	96.5% (96.1%-96.9%) 7,197/7,456
Specificity	94.9% (94.6%-95.2%) 20,461/21,550	89.6% (89.3%-89.9%) 44,665/49,850	97.3% (97.2%-97.5%) 42,803/43,988	99.5% (99.4%-99.5%) 73,670/74,061
Positive predictive value (precision)	98.2% (98.1%-98.3%) 58,872/59,961	85.5% (85.2%-85.9%) 30,663/35,848	96.8% (96.6%-97.0%) 35,638/36,823	94.8% (94.3%-95.3%) 7,197/7,588
Negative predictive value	94.9% (94.6%-95.2%) 20,461/21,556	97.8% (97.7%-97.9%) 44,665/45,669	95.8% (95.6%-96.0%) 42,803/44,694	99.6% (99.6%-99.7%) 73,670/73,929

A threshold of 0.5 was utilized in calculating accuracy, sensitivity, specificity, positive predicted value, and negative predicted value.

revision rate between the control and test periods, it is also important to mention the increase in the number of laboratory events detected during the test period. An individual human error during the release of results within the LIS was involved in 9 out of the 12 events identified by the laboratory during the test period. These individual errors resulted in incorrect results being released for 9 separate patient kidney stones. This one error caused the laboratory-identified revision rate to be 3x that of the control period. It is important to note that even with the significant increase in events due to the implementation of the AI quality assurance program, the overall event rate during the study period was still very low at 0.096% (78/81,517).

DISCUSSION

This study has demonstrated that by collaborating across several specialties, including laboratory medicine, data science, and biostatistics, effective AI-based quality assurance programs can be created and successfully implemented in practice. Additionally, the implementation of human-AI-augmented workflows can result in a significant increase in the identification of laboratory errors and thus improved patient safety. This finding is similar to the results recently reported by Bates et al.[14] Laboratory errors are known to significantly impact patient care as it is estimated that 60%-70% of all diagnoses are based on laboratory testing.[15] Research has determined that erroneous laboratory results have been reported to be 0.012%-0.6% of all reported tests.[15-17] This finding is concurrent with that of the present study, which exhibited an overall error detection rate of 0.096%. This rate included 62 errors detected by the AI quality program, 4 errors detected by the client, and 12 errors detected by the laboratory. The present study exhibited that by incorporating AI spectra identification into our clinical workflow, we could potentially reduce a significant proportion of the human error associated with this test through continued use of the AI algorithms. This immediate reduction in laboratory errors not only improves patient care and safety by ensuring correct kidney stone identification but also potentially reduces costs associated with erroneous laboratory results, including

TABLE 3. Model Performance by 4 Most Prevalent Stone Types for Initial Model Validation and Prospective QA Program Results

Stone classification	Model accuracy		autoPass accuracy	
	Concordance rate (95% CI)	Concordance rate pass QA (95% CI)	% autoPass	Concordance rate pass QA (95% CI)
Initial model validation (n=16,491)				
Calcium oxalate monohydrate	80.9% (80.2%-81.6%) 9,891/12,228	97.1% (96.7%-97.3%) 11,868/12,228	46.5% (5,689/12,228)	99.7% (99.5%-99.8%) 5,673/5,689
Calcium oxalate dihydrate	68.0% (66.9%-69.1%) 4,768/7,007	94.5% (93.9%-95.0%) 6,620/7,007	13.6% (953/7,007)	98.2% (97.2%-99.0%) 936/953
Calcium phosphate	63.3% (62.1%-64.4%) 4,426/6,997	92.3% (91.6%-92.9%) 6,455/6,997	12.2% (854/6,997)	96.7% (95.3%-97.8%) 826/854
Uric acid	83.8% (82.1%-85.5%) 1,564/1,866	94.0% (92.8%-95.0%) 1,754/1,866	53.5% (998/1,866)	99.9% (99.4%-100.0%) 997/998
Prospective QA program (N=81,517)				
Calcium oxalate monohydrate	70.5% (70.1%-70.9%) 42,010/59,581	94.5% (94.3%-94.7%) 56,307/59,581	43.0% (25,640/59,581)	99.2% (99.1%-99.3%) 25,442/25,640
Calcium oxalate dihydrate	57.7% (57.2%-58.3%) 18,120/31,391	90.6% (90.3%-91.0%) 28,454/31,391	12.3% (3,858/31,391)	97.1% (96.5%-97.6%) 3,746/3,858
Calcium phosphate	51.1% (50.6%-51.6%) 18,951/37,080	85.9% (85.5%-86.2%) 31,839/37,080	10.4% (3,838/37,080)	94.1% (93.3%-94.9%) 3,613/3,838
Uric acid	68.6% (67.6%-69.6%) 5,739/8,366	88.7% (88.0%-89.3%) 7,418/8,366	49.2% (4,113/8,366)	98.4% (97.9%-98.7%) 4,046/4,113

Concordance rate = number of true classifications (exact matches)/total spectra reviewed. Concordance rate pass QA = number of true classifications after application of quality assurance rules (comparable matches)/total spectra reviewed.

autoPass = automatic classification.

TABLE 4. Summary of Corrected Events for the Amended Kidney Stones Results in the Laboratory Information System

	Control period (n=82,294)		Test period (N=81,517)		
	N	Correction rate[a]	N	Correction rate[a]	RR (95% CI)
Overall	10	1.22	78	9.57	7.87 (4.08-15.2)
Lab	4	0.49	12	1.47	3.03 (0.98-9.39)
Client	6	0.73	4	0.49	0.67 (0.19-2.38)
AI	-	-	62	7.61	-

[a] Correction rate was summarized per 10,000 stones.

Lab = event identified by the laboratory; Client = event identified by treating physician's request for clarification; AI = event identified by the AI quality assurance program. Control period: October 29, 2019, to October 28, 2020. Test period: October 29, 2020, to October 28, 2021.

those associated with misdiagnosis, inappropriate treatments, and case management costs.[18]

Machine learning and deep learning techniques have been applied to identifying kidney stone composition by various methods and data sources. Abraham et al[19] used machine learning to predict kidney stone composition using electronic health record data including 24-hour urine testing data. Black et al[20] and Stone[21] used deep learning to assess kidney stone composition from digital photographs. Additionally, Cui et al[11] used Raman spectroscopy in combination with machine learning techniques to classify kidney stones. Although these studies often contained small samples sizes, they support the premise that the use of advanced statistical techniques can provide utility for the identification of kidney stone constituents. Important differences, however, exist between these approaches and the approaches taken with this study. Studies often classified kidney stones only by the major stone constituent, and accuracy of the various models was significantly diminished when kidney stones contained more than one constituent. The advantages of the present study included training and validation of the algorithm using a large, diverse, and clinically-derived kidney stone spectra data set. Furthermore, the present study used analytical instrumentation that is currently available and readily used in several clinical reference laboratories for kidney stone constituent analysis, which ensures the potential to generalize the approaches further.

The study took a conservative approach in that the AI results were generated asynchronously from the initial technician review. This was by design as the full implementation of the AI technologies was considered premature until the operating characteristics in a clinical setting were better understood. During the study, however, the reporting and the timeliness of the AI predictions were improved. This was in response to the value the AI predictions brought to the standard testing workflow and the need to mitigate errors reported to referring providers. In doing so, the present study brought AI autonomous reporting one step closer to the clinical laboratory. However, several barriers must be properly addressed before AI autonomous reporting of kidney stone FTIR spectra becomes a reality in the clinical laboratory. These include regulatory, information technology infrastructure and staffing barriers that currently limit the ability of AI algorithms from directly reporting results to the provider. Although scientific societies and regulatory bodies are beginning to create guidelines for the implementation of AI-based diagnostics, appropriate validation of new applications may not be well defined.[22] Additionally, the creation, modification, and storage of scalable AI algorithms requires significant changes to computing infrastructure. The computing infrastructure changes include remapping the data workflow between the spectrometer and the technician who reads the spectrum on the display monitor. Between these 2 points, an automated backend leveraging resting state application programming interface will be required to present the AI interpretations in real time to the technician. Finally, laboratory staff buy-in is critical in successfully implementing an AI-based reporting system into practice. As was learned in this quality assurance program, AI has the potential to augment, not replace, the trained technicians.

One of the design decisions for this research was recognition that a single algorithm to classify every possible FTIR spectra may not be sufficient or generally effective. Thus, a suite of algorithms was created and validated. The advantage of the suite of algorithms, particularly the constituent-specific algorithms, must be identified. To answer this question, consider a hypothetical scenario in which you want to develop a computer vision algorithm to identify the make, model, and year of vehicles driving on the interstate systems. One algorithm might be able to readily identify the make and models of cars. The year of manufacture may prove to be challenging for a computer vision algorithm, particularly if model years change and there was no retraining of the algorithms. Specific algorithms might be able to discern the color of the vehicle. Likewise, algorithms for the broad class of vehicle (e.g., sedan, pickup truck) would be technically feasible. In the event the single model classifier did not provide a reliable prediction, these supplemental models may provide guidance that can rapidly limit the number of candidate vehicles through providing separate predicted classifications for the make, model, and generation of vehicle. This is effectively how our constituent-specific models are configured: an aide to the many-class model if this model prediction did not pass the criteria to be automatically classified.

The present study has certain limitations. A major limitation of the quality assurance program was that the manner in which the AI results were used by the practice changed over time. These changes were not a reflection of changes to the models; those were fixed and stabilized at the start of the program. What changed is how the AI results would be reviewed and consumed in practice. Changes in the report content and frequency of generation evolved over the study period. This met the practical need of the laboratory staff, albeit at the expense of scientific purity. Another important limitation pertains to the training of the original models. In the training of an AI model, one generally assumes the labels are the truth. We observed some exceptions of this assumption during the validation study. Going forward, we will be utilizing the results of the quality assurance program to revisit the original training data to provide an improved training set of data. Further validation studies may be required to measure the performance of the newly trained versions of the algorithms.

CONCLUSION

This study demonstrated that it is possible to create and implement an AI quality assurance program that can accurately interpret FTIR kidney stone spectra, decrease laboratory errors, and ultimately lead to better patient management. Additionally, with the incorporation of an AI-augmented workflow within the clinical laboratory, the laboratory was able to identify and correctly report out the kidney stone constituents at a higher rate, which is crucial for treatment and recurrence prevention in the stone-forming patient.

POTENTIAL COMPETING INTEREST

Dr. Rickey Carter, an editorial board member, had no role in the editorial review of or decision to publish this article. All other authors report no competing interests.

SUPPLEMENTAL ONLINE MATERIAL

Supplemental material can be found online at https://www.mcpdigitalhealth.org/. Supplemental material attached to journal articles has not been edited, and the authors take responsibility for the accuracy of all data.

Abbreviations and Acronyms: **AI,** artificial intelligence; **FTIR,** Fourier transform infrared; **LEMS,** Laboratory Event Management System; **LIS,** laboratory information system; **SHAP,** SHapley Additive exPlanations

Grant Support: Portions of this work were supported by the Innovation Accelerator Awards from the Mayo Clinic Office of Translation to Practice. The content is solely the responsibility of the authors and does not necessarily represent the official views of Mayo Clinic.

Correspondence: Address to Mayo Clinic, 4500 San Pablo Road, Jacksonville, FL 32224 (Carter.rickey@mayo.edu).

ORCID

Rickey E. Carter: https://orcid.org/0000-0002-0818-273X

REFERENCES

1. Alelign T, Petros B. Kidney stone disease: an update on current concepts. *Adv Urol*. 2018;2018:3068365.
2. Romero V, Akpinar H, Assimos DG. Kidney stones: a global picture of prevalence, incidence, and associated risk factors. *Rev Urol*. 2010;12(2—3):e86-e96.
3. Hyams ES, Matlaga BR. Economic impact of urinary stones. *Transl Androl Urol*. 2014;3(3):278-283.
4. Strohmaier WL. Economics of stone disease/treatment. *Arab J Urol*. 2012;10(3):273-278.
5. Rule AD, Lieske JC, Li X, Melton LJ 3rd, Krambeck AE, Bergstralh EJ. The ROKS nomogram for predicting a second symptomatic stone episode. *J Am Soc Nephrol*. 2014;25(12): 2878-2886.
6. Gambaro G, Croppi E, Coe F, et al. Metabolic diagnosis and medical prevention of calcium nephrolithiasis and its systemic manifestations: a consensus statement. *J Nephrol*. 2016;29(6): 715-734.
7. Frassetto L, Kohlstadt I. Treatment and prevention of kidney stones: an update. *Am Fam Physician*. 2011;84(11):1234-1242.
8. Khan AH, Imran S, Talati J, Jafri L. Fourier transform infrared spectroscopy for analysis of kidney stones. *Investig Clin Urol*. 2018;59(1):32-37.
9. Choudhury A, Asan O. Role of artificial intelligence in patient safety outcomes: systematic literature review. *JMIR Med Inform*. 2020;8(7):e18599.
10. Paranjape K, Schinkel M, Hammer RD, et al. The value of artificial intelligence in laboratory medicine. *Am J Clin Pathol*. 2021; 155(6):823-831.
11. Cui M, Zhang DY. Artificial intelligence and computational pathology. *Lab Invest*. 2021;101(4):412-422.
12. Lundberg SM, Lee S-I. In: Guyon I, Luxburg UV, Bengio S, et al., eds. *A unified approach to interpreting model predictions*. 2017;30.
13. Kundu S. AI in medicine must be explainable. *Nat Med*. 2021; 27(8):1328.
14. Bates DW, Levine D, Syrowatka A, et al. The potential of artificial intelligence to improve patient safety: a scoping review. *NPJ Digit Med*. 2021;4:54.
15. Agarwal R. Quality-improvement measures as effective ways of preventing laboratory errors. *Lab Med*. 2014;45(2):e80-e88.
16. Plebani M, Carraro P. Mistakes in a stat laboratory: types and frequency. *Clin Chem*. 1997;43(8 Pt 1):1348-1351.
17. Carraro P, Plebani M. Errors in a stat laboratory: types and frequencies 10 years later. *Clin Chem*. 2007;53(7):1338-1342.
18. Northrup JM, Miller AC, Nardell E, et al. Estimated costs of false laboratory diagnoses of tuberculosis in three patients. *Emerg Infect Dis*. 2002;8(11):1264-1270.
19. Abraham A, Kavoussi NL, Sui W, Bejan C, Capra JA, Hsi R. Machine learning prediction of kidney stone composition using electronic health record-derived features. *J Endourol*. 2022; 36(2):243-250.
20. Black KM, Law H, Aldoukhi A, Deng J, Ghani KR. Deep learning computer vision algorithm for detecting kidney stone composition. *BJU Int*. 2020;125(6):920-924.
21. Stone L. Assessing kidney stone composition using deep learning. *Nat Rev Urol*. 2020;17(4):192-193.
22. Park SH, Choi J, Byeon JS. Key principles of clinical validation, device approval, and insurance coverage decisions of artificial intelligence. *Korean J Radiol*. 2021;22(3):442-453.